1. Introduction

Deep learning has been demonstrated as a unified learning paradigm capable of modeling and predicting new types of relationships in multiple images. We establish an evaluation protocol that properly pairs image regions with text phrases in certain relationships, object parts, and objects with part relationships. We suppose the regions of interest may be different across images.

2. Discriminative Bimodal Networks (DBNet)

- Text features: Character-level CNN
- Replicating a phrase until reaching the input length of the CNN (256 characters)
- Non-saturating activations: Leaky ReLU
- Can be replaced by other text embedding techniques

The detection score of image region $x$ is given by a linear classifier dynamically generated according to the text phrase $t$:

$$ s_x(t) = w(t) \cdot \phi_x(x; r, t) + b $$

where the classifier weights and bias are $w(t) = A^{-1} \sum_{x} \phi_x(x; r, t)$ and $b = A^{-1} \sum_{x} b_x(x; r, t)$.

An extra regularization term on the classifier parameter is important for the SGD stability:

$$ \gamma_{reg} = \frac{1}{2} \sum_{x} ||w(t)||^2 $$

3. Model Architectures

- Based on the Visual Genome dataset
- Bounding boxes with text phrase descriptions
- Additional spell checking and auto-correction
- Localization: finds a known-to-exists entity
- Detection: localize all matched entities if exists any
- Should have negative images (no matched entity)
- Impractical to test all queries on each image

The first benchmark protocol for visual entity detection with language queries. 3 difficulty levels with increasing number of randomly chosen negative images per query:

- Level 0: no negative image
- Level 1: the same number as the positive image
- Level 2: 5 times the number of positive images or 20 (whichever is larger) for each test phrase

4. Model Training: Labels, Objectives, and Optimization

- Any region-text pair $(r, t)$ can be possibly used in training.
- $r$ can be GT or proposed (EdgeBox), $t$ can be an annotation on the same image as $r$ or from the rest of the training set.

- Spatial overlapping based training labels:
  - $l_i$: has large overlap with a ground truth region of $t$
  - Uncertainty: $(r, t)$ is not positive, and $r$ has moderate overlap with a ground truth region of $t$ (excluded from training)

- Text similarity based uncertainty augmentation:
  - If $l_i$ is similar to $l_j$, $l_j$ should also have uncertain label with $r$
  - $r$: otherwise (including all other text phrases).

- Given a region, the non-uncertain phrases are categorized into: positive phrases (pos), negative phrases from the annotations on the same image (neg), and negative phrases from the rest images (noc).

- Training loss is normalized separately for the three categories:

$$ L = \sum_{i} \left( \lambda_{pos} \max_{t} s_i(t) - \lambda_{neg} \min_{t} s_i(t) - \lambda_{noc} - \eta \right) $$

where we set $\lambda_{pos} = \lambda_{neg} + \lambda_{noc}$, the sampling strategy in SGD determines $\lambda_o$ (larger) and $\lambda_o$ (smaller).

- Optimization: (initialization) Initializing visual pathway with pretrained faster R-CNN;
  - (phase 1) training the text pathway from scratch and fix the visual pathway parameters;
  - (phase 2) jointly finetuning the two pathways: (phase 3) decrease the learning rate.

5. Benchmarking Protocol

- Accuracy/%
- Difficulty level:

6. Experiments

We release our implementation in both Caffe/MATLAB (original results) and TensorFlow.

- DBNet training is memory consuming. Our code perform subbatch partition for subnetworks to fit the GPU memory limit while keeping a large overall batch size for effective training.

- For an English text $t$, we encode each of its characters into a 74-dim character encodings. We use a character-level deep CNN in the character-level CNN.

- A $f_{txt}$ extracts the image features $v_{pos}$ which are assigned with the "uncertain" label to avoid false negatives. Any $v_{neg}$ can be GT or proposed (EdgeBox) when applicable. DBNet outperformed DenseCap and SCRC significantly. DenseCap and SCRC show higher per character performance in term of MAP.

- DBNet’s scores are better “calibrated” over different phrases in terms of gAP.

- DBNet shows more robustness to negative images and rare text phrases.

7. Comparative Quantitative

- Ablation study of DBNet’s major components

8. References